Treatment, prediction, and assessment of childhood aggression

Hendriks, A.M.

2019

Chapter 4

Predicting Childhood Aggression: Mining Large Datasets Followed by Confirmatory Modeling.

Submitted as: Hendriks, A. M., Luningham, J., Hong, M., Jaccobucci, R., Lundström, S., Larsson, H., … Lubke, G. (2019). Predicting childhood aggression: Mining large data followed by confirmatory models.

ABSTRACT

Background: The aim was to predict childhood aggression, by analyzing data from two large European cohorts (N = 62,227), with a novel methodological approach.

Data: Participants came from the Child and Adolescent Twins Study in Sweden and the Netherlands Twin Register. The outcome measure was physical overt aggression as assessed around age 9, psychometrically harmonized across multiple European cohorts. The large set of predictor variables consisted of demographics, prenatal characteristics, physical development, family environment, parenting, parental education level, life events, and behavioral symptoms.

Method: To avoid capitalization of chance, data were partitioned in four parts for different analysis steps. These included 1) exploratory data analysis and tuning meta-parameters for data mining, 2) fitting increasingly complex data mining models to assess which predictors had which types of effects, 3) assessment of model performance and importance of the predictor variables, and 4) fitting a confirmatory prediction model of aggression that integrated results of the data mining analyses.

Results: The prediction model confirmed linear main effects of predictor variables and included interactions of predictors with sex and cohort. Associations between the main predictors (non-physical aggression, ADHD, conduct disorder, maternal smoking during pregnancy, parenting style, and life events) and childhood aggression were in line with previous research, yet weaker, likely because we considered more predictors simultaneously.

Conclusion: Fitting all predictors simultaneously provided clear insight in the importance of predictors relative to each other. Externalizing, non-aggressive behaviors had the strongest effects, and may act as salient targets for early detection and prevention of childhood aggression.

Keywords:
Childhood aggression, data mining, confirmatory prediction model

GENERAL SCIENTIFIC SUMMARY

Salient behaviors associated with ODD, CD, and ADHD such as arguing, being easily distracted, and hyperactivity appear to best predict childhood aggression, above prenatal, physical, and environmental predictors (e.g., maternal smoking during pregnancy, parental education level). This is consistent with previously reported high comorbidity of childhood aggression with other behavioral and emotional problems. These behaviors, which could be noticed by people in the environment, may facilitate early detection and prevention of (behavior problems related to) childhood aggression.

Although childhood aggression has been the focus of a large body of research, insight in the mechanisms of predictors on the development of childhood aggression and its psychosocial factors is still limited (Jaffee, Strait, & Odgers, 2012). Childhood aggression receives much attention due to its high prevalence (NICE, 2013), comorbidity with other disorders (Bartels et al., 2018), adverse outcomes for the individual (Copeland, Wolke, Shanahan, & Costello, 2015), negative consequences for parents (Meltzer, Ford, Goodman, & Vostanis, 2011; Roberts, McCrory, Joffe, De Lima, & Viding, 2018), and high costs for society (Rivenbark et al., 2018; Romeo, Knapp, & Scott, 2006). Although there is a small decrease in the prevalence over the past years (Erskine et al., 2014; Pickett et al., 2013), current treatment effects are still generally small (Hendriks, Bartels, Colins, & Finkenauer, 2018; Weisz et al., 2017). Given the burden on the individual and her/his surroundings as well as society at large, further research is needed to increase knowledge about the precursors of aggression. The ACTION (Aggression in Children: Unraveling gene-environment interplay to inform Treatment and interventION strategies) consortium aims to combine multidisciplinary information from multiple research groups to enhance knowledge on childhood aggression (Bartels et al., 2018; Boomsma, 2015). The present study utilizes the combined sample size and wealth of information in the ACTION consortium data to find a set of robust predictors for childhood aggression, to assess their respective importance using advanced analysis, and to describe their relation to childhood aggression.

The cohorts within the ACTION consortium contain a wide range of variables, which allowed us to include a heterogeneous set of variables to predict childhood aggression. Variable categories comprise of demographics, prenatal characteristics, physical development, the family environment, parenting, parental education level, life events, and behavioral symptoms as reported by mothers. For example, pre- and perinatal characteristics include variables such as birth weight or maternal smoking during pregnancy. Behavioral symptoms as reported by the mother may refer to behaviors associated with forms of childhood psychopathology (e.g., attention-deficit/hyperactivity disorder, anxiety).

While multiple studies have examined some of these predictors of childhood aggression, most focused on individual predictor variables, for example, sex of the child (Archer, 2004; Card, Stucky, Sawalani, & Little, 2008) or socioeconomic status (Piotrowska, Stride, Croft, & Rowe, 2015). Assessing predictors separately may however result in an incorrect estimation of predictor effects because predictors are correlated. Examining predictor variables simultaneously provides unbiased estimates as well as information concerning the importance of predictor variables relative to others (e.g., Sabina & Banyard, 2015). The present study combines two data-rich cohorts in order to investigate which of the different available predictor variables are associated with childhood aggression through main and/or interaction effects. The combined sample size in the present study, however, permits not only investigating large numbers of predictors and assessing potential nonlinear and interaction effects using data mining, but also to follow up with confirmatory analyses in a hold-out set of the data.

Data mining techniques are ideal for handling large data with many variables because there is no need to a priori specify what type of effect (linear/non-linear, main effect/interaction) a given predictor has on the outcome (Miller, Lubke, McArtor, & Bergeman, 2016). This is an important advantage since it is usually impossible to estimate a confirmatory model that includes main and interaction effects of all potential predictors. In this study we used an approach termed "Deductive Data Mining" (DDM; Hong et al., submitted) to inform which effects and variables need to be included in a final confirmatory model to predict childhood aggression. In DDM, increasingly complex data mining models are fitted to the data that differ with respect to the type of permitted effect a predictor can have on the outcome. For instance, the lasso (Tibshirani, 1996) only fits linear main effects whereas tree methods permit nonlinear and interaction effects. By comparing the performance of the different data-mining models one can deduce which predictors and which types of effects lead to the best model performance. Subsequently, the found effects are included in a confirmatory model that is fitted to a hold-out set of the data (i.e., a part of the data that has not previously been used for modeling; e.g., Faraway, 2016) in order to estimate effect sizes and perform statistical significance testing.

In addition to executing a larger scale search for potential predictors of childhood aggression, our study also combined two large European cohorts. Combining data sets from multi-county cohorts increases generalizability, but also poses methodological challenges because predictors and outcomes are often assessed by different instruments. Different item wording can introduce bias in parameter estimates when fitting models to the combined data, which in turn complicates interpretation (Curran & Hussong, 2009; Curran et al., 2008). Within the ACTION consortium, a physical aggression phenotype was harmonized using psychometric modeling (Luningham et al., submitted). This harmonized phenotype served as the outcome in the present study. Regarding the predictor variables, psychometric harmonization was not feasible due to the lack of a phenotypic reference set (e.g., a subsample with overlapping data on all questionnaires). Therefore, we followed the general practice of aligning items based on inspection of item wording followed by thorough exploratory analyses to detect potential cohort differences.

In conclusion, the goal of the present study was to find a set of robust predictors of childhood aggression and investigate how they relate to physical/overt childhood aggression using data from two large ACTION consortium cohorts. The present study provides a significant methodological innovation as this is a first large data mining study followed by a confirmatory analysis applied to substantive data from multiple cohorts, thus permitting recommendations for future collaborative projects. The paper is organized as follows. After describing the data resources, we provide an overview of the analysis strategy followed by a detailed explanation of each analysis step. This part includes the rationale of DDM and information concerning the specific data mining models as well as the measures to evaluate their performance. The results are presented accordingly. We finish by discussing the clinical conclusions and the limitations of our study.

METHOD

Data

The ACTION consortium comprises multiple large general population data sets from different cohorts, most of which assessed childhood aggression differently. To facilitate multi-cohort analyses within the ACTION consortium, a harmonized aggression score was created of childhood aggression scores through psychometric modeling (Luningham et al., submitted). The harmonization of the outcome variable was carried out using data from the Child and Adolescent Twins Study in Sweden (CATSS, N = 27,894; Anckarsäter et al., 2011), FinnTwin12 (FT12, N = 4,884; Kaprio, 2013), the Netherlands Twin Register (NTR, N = 34,333; Van Beijsterveldt et al., 2013), The Swedish Twin study of CHild and Adolescent Development (TCHAD, N = 2,181; Lichtenstein, Tuvblad, Larsson, & Carlström, 2007), and the Twins Early Development Study (TEDS, N = 17,267; Haworth, Davis, & Plomin, 2013). The overlap of data available as predictors for aggression across all cohorts was small, with CATSS and NTR forming the exception. CATSS and NTR are the two largest ACTION cohorts (total N = 62,227), and had 27 comparable items tapping into the domains of child and parent characteristics as well as mother-rated ADHD indicators. All analyses were carried out in these two cohorts.

CATSS. CATSS was launched in 2004 in order to longitudinally follow development of Swedish twins born in Sweden since 1992 during childhood and adolescence. First data collection was through a telephone interview with the parents of 9/12 year-old twins, followed by questionnaires at ages 9, 12, 15, and 18 (Anckarsäter et al., 2011). The sample for which the harmonized aggression score was available consisted of 27,894 CATSS participants.

NTR. NTR is a nationwide population-based register founded in 1987 in the Netherlands to investigate individual differences in mental and physical health. Data collection starts with a questionnaire shortly after birth of the twins, which is followed by age-specific questionnaires at age 2, 3, 5, 7, 9/10, 14, 16, and 18 (Van Beijsterveldt et al., 2013). The sample for which the harmonized aggression score was available consisted of 34,333 NTR participants.

Variables

The outcome: A harmonized factor score of overt/physical aggression. The five ACTION cohorts employed different questionnaires to assess childhood aggression, namely the Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001), the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997, 2001) and the Autism - tics, attention-deficit hyperactivity disorder and other comorbidities (A-TAC; Hansson, Svanstrom Rojvall, Rastam, Gillberg, & Anckarsater, 2005). Prior to the present study, the aggression phenotype was harmonized across five major participating cohorts of ACTION. Items pertaining to overt/physical aggression were assessed in each of the different questionnaires and combined to model a harmonized aggression score (items listed in Supplementary Table 1). In order to facilitate psychometric modeling additional data were collected on a subsample of 9-year old twins in the NTR (N = 2,316; 2,263 twins with mother report and 1,548 with father report). Mothers and fathers rated their children on all questionnaire items used in ACTION cohorts, resulting in a so-called “reference set”. The reference set permitted modeling item level data and the extraction of a factor score of childhood aggression while controlling for sex, rater, and cohort differences. This factor score represented a generalizable measure of overt or physical aggression (Luningham et al., submitted)

The predictors. Questionnaire items were selected with similar content across the two cohorts. We obtained variables in the following categories: demographics, prenatal characteristics, physical development, family environment, parenting, parental education level, life events, and behavioral symptoms as reported by the mother. Demographics comprised of sex, age of the child at the assessment of aggression, and cohort (i.e., CATSS, NTR). Pre- and perinatal characteristics were birth weight, gestational age, maternal smoking during pregnancy, and maternal alcohol use during pregnancy and were assessed at age 9/12 for CATSS and at first contact for the NTR. For CATSS, data collection began at 9/12, so most of the predictor variables were collected at this age. For NTR, data were collected from shortly after birth up to 18 years, but because the sample for age 7 years was largest, we selected this cohort for most variables. Physical development variables consisted of height, weight, asthma, eczema, and medication use. The category family environment included whether a child had siblings (not present in CATSS data), age of mother at birth, age of father at birth, and whether both parents lived in the same household. Parenting assessed parental monitoring. Parental education level comprised of maternal and paternal education level. Life events referred to the proportion of life events that children experienced; both CATSS and the NTR included a list of life events, however containing a different number of items and with different content. To harmonize this variable, we calculated the proportion of life events to which the response was “yes” out of the total number of life events. Behavioral symptoms as reported by the mother (assessed at age 9/12 in CATSS and age 7 in NTR) consisted of motor skills, arguing, lying, bragging, feeling no guilt, short attention, daydreaming, easily distracted, not finishing things, hyperactivity, feeling superior, being bullied, impulsivity, fear of situations, anxiety, and nightmares. Supplementary Table 2 presents information on all variables, whether they were available across both cohorts, and exact item wording for the variables with more than just minor differences in wording.

Because children in the NTR were around age 7 at the included assessment of height and weight and children in CATSS were either age 9 or 12, we corrected height based on an average difference of 11.30 cm between age 7 and 9, and an average difference of 17.60 cm between age 9 and 12 (Bonthuis et al., 2012). Similarly, for weight we corrected for a difference of 5.25 kg between age 7 and 9 (World Health Organization, 2007). Because of growth spurts after age 10, however, the WHO did not report average weight beyond this age, so we corrected for the mean difference in the CATSS data between age 9 and 12 (11.40 kg), which was in line with growth numbers reported for the United Kingdom (Royal College of Pediatrics and Child Health, 2012). Unrealistic values for height (i.e., below 105 cm, above 145 cm) and weight (i.e., below 15 kg, above 45 kg) were considered as missing data (World Health Organization, 2007).

The items related to the quality of parenting differed for CATSS and NTR. NTR started to include these variables in 2010, resulting in a high missingness rate (88%) in the current sample. We therefore tested the effects in cohort specific analyses because of their theoretical importance (e.g., Racz & McMahon, 2011; Stattin & Kerr, 2000). Similar as for parenting, the life events items varied across cohorts, and were thus included in cohort specific analyses (e.g., Grant, Compas, Thurm, McMahon, & Gipson, 2004; Guerra, Huesmann, Tolan, Van Acker, & Eron, 1995).

The behavioral symptoms in the CATSS data were from the A-TAC, with response options 0 = *"No"*, 0.5 = *"Yes, to some extent"*, and 1 = *"Yes"*. The behavioral symptoms in the NTR were from the CBCL, with response options 0 = *"Not true"*, 1 = *"Somewhat or sometimes true"*, and 2 = *"Very true or often true"*. Although the items were measured on 3-point scales in both cohorts, response options were dichotomized due to a differential use of the zero and mid point response options in the two cohorts, with CATSS having a higher frequency of zero than the NTR whereas the pattern was reverse for the mid category. Dichotomization was done by collapsing the zero and midpoint response options in both cohorts, which resulted in very similar item distributions across questionnaires.

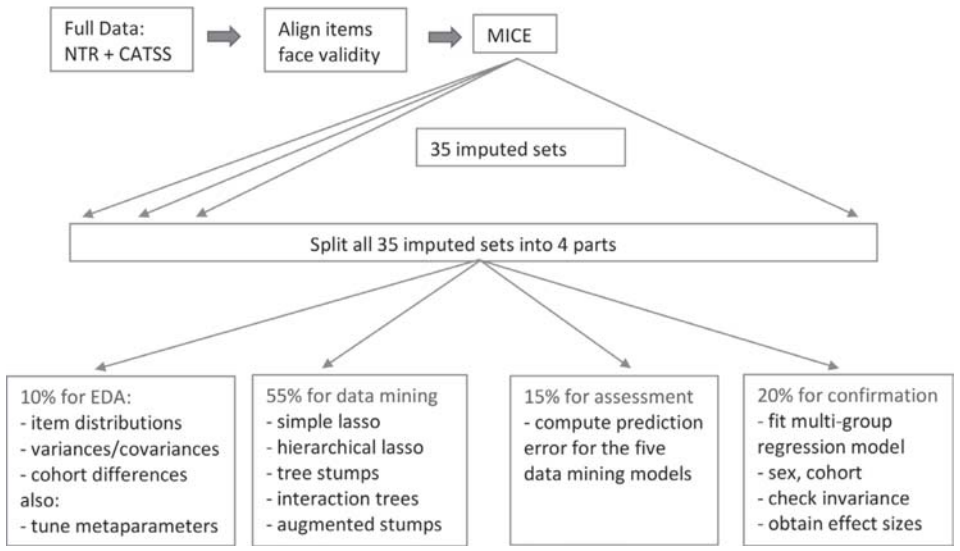
Table 1. Workflow of the analyses

Part	Percentage	Analysis step
	100%	Imputation of missing data - Method was multiple imputation by chained equations (MICE) - It creates multiple imputed data sets over which model outcomes may be averaged to obtain more robust results and insight in the variability of outcomes - followed by splitting the data into 4 parts
1	10%	Exploratory data analysis - Structural stratification in the predictor variables. E.g., cohort differences, sex differences - Check predictor variables for near-zero variances - Examine whether there are high correlations (> .70) among predictor variables - Tune meta-parameters for the data mining analyses
2	55%	Data mining - Simple lasso for linear main effects - Hierarchical lasso for linear main and interaction effects - Boosted tree stumps for linear and non-linear main effects - Boosted interaction trees for linear and non-linear main and interaction effects
3	15%	Model performance assessment - Use prediction error and proportion of variance explained as a measure of model performance for the different data mining methods - Variable importance measures to assess the relative importance of each predictor in predicting childhood aggression - Look for convergence across models
4	20%	Confirmatory prediction model - Fit a model based on the outcomes of the data mining analyses to predict childhood aggression to obtain effect sizes of the relations between the predictor variables and childhood aggression

Analysis Strategy

The analysis of the ACTION data is outlined here and described in detail in the next paragraphs. The workflow is shown in Table 1 and graphically displayed in Figure 1. Prior to the main analyses we carried out multiple imputation by chained equations (MICE to impute missing data, resulting in 35 imputed sets. Next, each of the imputed sets was split into four mutually exclusive parts. This is necessary to avoid capitalization of chance and inflation of Type I error that occurs in sequential analyses without data splitting (Hastie, Tibshirani, & Friedman, 2009; Lubke & Campbell, 2016). Steps 1-4 of the main analyses were carried out in the four different parts of the data. Step 1 involved Exploratory Data Analysis checks and selection of meta-parameters for the data mining models (Part 1 of the data comprised 10% of the sample). The 2nd analysis step implemented the deductive data mining (DDM) procedure using Part 2 (55% of the data). The performance of the different models was assessed in Part 3 of the data (step 3, 15% of the data). The 4th and final step consisted of integrating the results from steps 2 and 3 in a traditional confirmatory model prediction aggression fitted to Part 4 of the data (20%).

Figure 1. Process flow



Step 1: Imputation.

Multiple imputation is a well-known and powerful approach for dealing with complex missing data. In multiple imputation, missing values are filled in by creating a prediction model for the missingness and drawing values from the predictive distribution repeatedly to create multiple complete datasets (Rubin, 1987; Van Buuren & Groothuis-Oudshoorn, 2011). The analysis of interest is carried out in each of the complete datasets and subsequently pooled (Rubin, 1987). The imputation process is more challenging when missing data are present in multiple variables within a dataset. Multivariate missing data can be handled by MICE, in which an iterative process is set up establishing a predictive distribution for each variable with missingness conditional on all other variables (Van Buuren & Groothuis-Oudshoorn, 2011). The procedure is as follows. For the first variable with missing values,

a predictive distribution is established by predicting the variable from all other variables in the dataset for cases in which the variable was not missing. Missing values are filled in by draws from this predictive distribution. The next variable with missing values is then predicted and filled in conditional on all other variables in the data. One iteration is complete after cycling through all variables with missing values. The process continues for multiple iterations to create an imputed dataset, and this is repeated to build multiple imputed datasets.

MICE is advantageous because of its flexibility and its utility with large, complex data. In MICE, the researcher can utilize different models for each variable to establish the predictive distribution; for example, a linear regression can be specified for a continuous variable such as age and a multinomial regression used for a categorical survey item. Additionally, creating multiple imputations reflects uncertainty in the prediction of missing data. Predictive mean matching (PMM) is a common imputation procedure that introduces missing data uncertainty into the imputation. PMM utilizes a linear model to predict a variable with missingness from the other variables, but it includes two steps to add randomness to the procedure. In each iteration, the regression coefficients are first drawn from a multivariate normal distribution based on their estimates and covariance matrix. Then, a predicted value is found for cases in which the variable is missing. However, the imputed value is not the predicted value from the linear model, but is instead a random draw of observed values that are nearest to the predicted value. The number of candidate “neighbors” is typically around 5, 10, or 15. These two steps prevent bias due to overconfidence of the predicted values for missing data.

For the current study, nearly all variables of interest contained some missing data, with the exception of the harmonized aggression score. Multiple imputation was used to obtain the largest number of complete cases possible for the variable selection step, as some of the chosen data mining algorithms rely on complete cases. However, there is evidence that imputing variables with extreme levels of missingness leads to bias and lower power, so we only imputed variables with reasonable amounts of missing data (Kontopantelis, White, Sperrin, & Buchan, 2017). Variables were not included in the imputation procedure if a) the variable had more than 50% missing values in the data combined across NTR and CATSS, or b) the variable was 100% missing within either NTR or CATSS. In all, 26 covariates were imputed prior to data-splitting and subsequent analyses. Each covariate was imputed from all other covariates that met the missing proportion threshold for imputation plus the harmonized aggression outcome, consistent with recommended practices (Sterne et al., 2009; Van Buuren & Groothuis-Oudshoorn, 2011). Imputation was implemented using the R package mice (Van Buuren & Groothuis-Oudshoorn, 2011).

Step 2: Partitioning of the data.

The motivation to split the data and carry out the different steps of the analysis in separate parts of the data is to prevent capitalization on chance, which occurs if sequential steps of an analysis are carried out in the same data. For instance, assessing the performance of a model in the same data that were used to fit the model would result in overestimating the performance. This is because a model adapts not only to the structure in the data but also to some extent to the idiosyncrasies of the sample. An evaluation of a model in new data shows how well the model can predict the actual structure.

We chose to partition the data into parts of different sizes to optimally leverage the total sample. The partitioning was carried out in the same way in all 35 imputed sets, and was based on random selection at the subject level. A different option would have been to select at the family level, which would have the advantage of creating independent partitions. However, currently the data mining methods used in this study do not offer to account for family relatedness. Since the largest part of the data is used for mining, we preferred to minimize the relatedness within partition by selecting at the subject level. A partition of 10% (N = 6,222) was sufficient to carry out exploratory data checks, and to tune the data mining algorithms. The main part of the analysis focused on fitting different types of data mining approaches. Due to fact that the low signal to noise ratio in behavioral data requires large samples (e.g., Ivanova et al., 2007), we choose to allocate 55% of the data (N= 34,225) for this step. Prediction error served to compare the different data mining models and was calculated in 15% (N = 9,334) of the data. Here we also computed variable importance measures and selected important predictors for the last step. The remaining 20% (N = 12,446) were used to fit a confirmatory model and assess statistical significance and effect sizes of the selected predictors.

Step 3: Exploratory Data Analysis and Meta-parameters.

Predictor variables were selected based on item content. Since the items stemmed from different questionnaires, the item wording was not identical. Therefore, exploratory analyses were necessary to check whether there was structural stratification across cohorts. In addition, we assessed whether variables had extremely low endorsement rates leading to near-zero variances, and whether there were high correlations between predictor variables (i.e., above .7). We then used the R package caret (Kuhn, 2018) to tune all data mining procedures to obtain optimal meta-parameters. The drawback of using a partition of the data that is smaller compared to the one used the main analysis is the potential dependence of meta-parameters on sample size. Therefore, a small subset of the meta-parameters (but not the entire grid) was checked again during the main analysis.

Step 4: Data Mining

In this analysis we used an approach termed "Deductive Data Mining" (DDM; Hong et al., submitted). Data mining methods in the regression setting are generally geared towards predictor selection and/or obtaining an optimal prediction performance (i.e., lowest prediction error in future data). Individual methods do not provide guidance regarding the type of effect of the predictors (e.g., linear/non-linear main and/or interaction effects). Deductive Data Mining (DDM) introduces the concept of model comparison that is well-known in confirmatory analyses such as Structural Equation Modeling into the area of data mining. The rationale of DDM is to compare the performance of data mining models that differ with respect to the type of effect that is included in the model. For instance, a simple lasso fits only linear main effects but no non-linear effects or interactions, whereas boosted trees adapt to linear as well as non-linear effects, and can include linear and non-linear interaction effects. By comparing the prediction error of a simple lasso and boosted trees one can assess whether the inclusion of non-linear and/or interaction effects is necessary to improve prediction performance. Table 2 provides an overview of the characteristics of the approaches used in this analysis, and the type of information they can provide in terms of effects. All methods afford the means to rank the predictors according to their importance of predicting aggression, and therefore permit variable selection.

In order of increasing complexity of effects, the different data mining approaches were (1) simple lasso (linear main effects), (2) hierarchical lasso (linear main and linear interaction effects), (3) boosted tree stumps (i.e., trees with a single split, linear and non-linear main effects), and (4) boosted interaction trees (linear and non-linear main and interaction effects). In the next section these four models are explained in more detail.

Table 2. Model descriptions for deductive data mining approach

Models	Linear marginal effects	Non-linear marginal effects	Linear interaction effects	Non-linear interactions
1) Lasso	All	N/A	N/A	N/A
2) Hierarchical lasso	All	N/A	All	N/A
3) Boosted stump tree (tree depth = 1)	All	All	N/A	N/A
4) Boosted interaction tree (tree depth = 5)	All	All	All	All
5) Confirmatory regression model with new data	Specified	Specified	Specified	Specified

Regularization methods: Lasso and hierarchical lasso

The least absolute shrinkage and selection operator (*lasso;* Tibshirani, 1996, 2011) as well as the hierarchical lasso (Bien, Taylor, & Tibshirani, 2013; Choi, Li, & Zhu, 2010; Haris, Witten, & Simon, 2016) are multiple regression models where regression coefficients are regularized. Regularization involves placing constraints (penalties) on the coefficients in a regression model such that the sum of square coefficients cannot exceed a penalty value, thus shrinking the coefficients from their OLS values. The aim of regularization is to improve the stability and the generalizability of the model. The lasso can shrink the coefficients all the way to 0, thereby performing variable selection.

Whereas the simple lasso only fits linear main effects, the *hierarchical lasso* includes interactions between variables and quadratic terms for each variable that is identified as a non-zero main effect. This produces a computationally efficient method to identify interactions and quadratic effects in a regression framework. Both lasso and hierarchical lasso only model linear effects. By comparing their performance to that of boosted stumps and interaction trees, respectively, one can deduct whether non-linear effects are necessary.

The amount of regularization in the lasso and hierarchical lasso is a meta-parameter that in our study was optimized in Part 1 of the data. The R package glmnet (Friedman et al., 2018; R Core Team, 2018) was used for the lasso to compare the fit of 100 penalty values using the lowest value of MSE to select the meta-parameter. A number of packages in R exist to implement the hierarchical lasso. We used the hierNet package (Bien & Tibshirani, 2014) to compare 20 penalty values, again with MSE as the selection criterion.

Boosted tree stumps and interaction trees

Trees are built in a recursive fashion (Breiman, 1984; Friedman, 2001). Similar to step-wise regression, the first step is to select the predictor that has the strongest association with the outcome. Rather than estimating a regression coefficient, a cut point on the predictor is obtained that optimally partitions the sample into two groups that are more homogeneous with respect to the outcome. For instance, suppose age is the strongest predictor of income. The algorithm searches for the optimal cut point (e.g., 25 years) that results in two groups (older/younger than 25) that are jointly the most homogenous with respect to income. The algorithm is then repeated in both partitions (called daughter nodes). The recursive partitioning of the nodes results in a tree structure. Single decision trees are popular because they are easy to interpret and visualize. However, the structure of a single tree structure depends heavily on the sample data. A new sample can result in a different choice of the splitting variables, and therefore in a very different tree structure.

Boosted trees combine a large number of trees to improve the prediction quality of a single decision tree (Friedman, 2001). The individual trees in such an ensemble can be specified to feature only a single split, resulting in so-called stumps. Since there are no conditional splits, tree stumps only capture main effects. Since many tree stumps are combined that may feature the same predictor but different cut points, stumps adapt to linear and non-linear main effects. Interaction effects are captured by permitting subsequent conditional splits. By comparing boosted stumps to trees with more splits (called interaction trees in this paper), one can deduct whether interactions are contributing to the prediction performance.

Trees can form the basis to rank all predictors according to their importance in predicting the outcome, but these methods do not lend themselves to deduct which of the individual variables are involved in the interactions. In order to evaluate specific second order interactions, we computed pairwise product terms of the predictors in each of the 35 imputed sets. These product terms were then included in augmented data sets to which we fitted tree stumps. The resulting variable importance measures provide the sought-after indication of which second order interactions are associated with the outcome. If the augmented stump model does not underperform the full-fledged boosted interaction model then one can deduct that only second order interactions are required.

Boosted trees require specification of three meta-parameters. The shrinkage parameter controls the contribution of each tree to the model, and therefore controls the speed with which the model adapts to the data. The shrinkage parameter is interrelated with the second meta-parameter, the number of trees to include in the model, with slower adaptation requiring more trees to be added to the model. The third parameter is the number of splits (also called tree depth), which controls the maximum order of interactions that can be modeled. In the current study we use the R package caret (Kuhn, 2018) to select these meta-parameters in Part 1 of the data and the package gbm (Greenwell, Boehmke, Cunningham, & GBM Developers, 2019, version number 2.1.5) to fit boosted trees to the aggression data.

Data mining analyses

We fitted the lasso, hierarchical lasso, boosted stump, and boosted interaction trees to part 2 of each of the 35 imputed data sets. Several potentially interesting predictors were only available in the NTR sample, and were investigated in additional analyses. These variables were (1) maternal alcohol use during pregnancy, (2) bragging, (3) feeling no guilt, (4) feeling superior, and (5) impulsivity. Items relating to parenting quality and the proportion of serious life events were only available in a small number of subjects in the NTR and were measured rather differently in the Dutch and Swedish samples. Therefore these items were only investigated in CATSS.

Step 5: Prediction Error and Variable Importance Measures

The performance of the four data mining models was assessed in part 3 of the data. We computed mean squared error (MSE) to quantify the prediction error of each model in new data. In addition, we obtained variable importance measures (VIMs; Friedman, 2001), which provide some guidance regarding which variables explain the most variance in childhood aggression.

Step 6: Confirmatory Model Predicting Overt/Physical Aggression in Children

Based on the comparison of the performance of the different data mining models it is possible to deduct which type of effects need to be included in a confirmatory model for an optimal prediction of childhood aggression. In the last part of the analysis we fitted a multi-group multiple regression model to part 4 of the data. The model included the effects of variables that appeared to be associated with childhood aggression model in deductive data mining.

RESULTS

Imputation

The imputation procedure was carried out with 30 iterations per dataset and 35 imputations in total. Thirty iterations were chosen to ensure proper convergence of the imputation for each dataset, though fewer iterations are typically required (e.g., 10-20; Van Buuren & Groothuis-Oudshoorn, 2011). Thirty-five datasets were imputed because the general recommendation is to create a number of datasets roughly equal to the percentage of missing data (Bodner, 2008). The average marginal missingness for the 26 imputed variables was 26.9%, so 35 imputations with 30 iterations each were expected to obtain a sufficient representation of the missing data. For all imputations, variability was introduced into the regression coefficients as described above. For continuous data, imputed values were drawn based on the 10 nearest neighbors. For binary outcomes, imputed values were found by drawing from a Bernoulli distribution using the predicted response probability.

All variables in the imputation were found to demonstrate good consistency and levels of convergence. The behavioral symptom variables generally reached a stable predicted value and standard deviation after around 20 iterations in each imputation, and the other variables reached convergence after about five iterations (convergence plots are presented in Supplementary Figure 1). The variables

generally demonstrated a small amount of between-imputation variability, especially the binary variables. This is likely due to the preponderance of zero's found in complete data, leading to a high likelihood of the imputed value also being zero.

Each imputed dataset was partitioned into the four analysis partitions. The imputed data were then checked by calculating descriptive statistics in the exploratory partition. Table 3 presents means, between- and within-imputation variability, and whether the variable was imputed based on the criteria of less than 50% missing values in combined data or 100% missing within a single cohort. Additionally, the proportion of missingness for each variable is presented in the combined data and within cohort.

Results Part 1 Data Analyses: Exploratory Data Analysis and Meta-Parameters

Low variances in the combined NTR/CATSS sample were detected for the following dichotomous items: motor skills (0.0262), lying (0.0142), stealing (0.0083), daydreaming (0.0524), being bullied (0.0214), fear of situations (0.0443), anxious (0.099), and nightmares (0.0157). Given the sample sizes in parts 2-4 of the analyses, these variances were deemed unproblematic as they translate to a sufficient number of endorsements in all parts of the data. Low variances in the NTR and CATSS cohort specific samples were very similar to the combined sample. A check for multicollinearity revealed that none of the correlations between predictors exceeded 0.75. Correlations with cohort that exceeded 0.1 were height (r=0.289, Dutch children taller), living with both parents (r=0.11, Dutch children more likely), eczema (r=.254, Swedish more likely), as well as parenting and life events. The latter two variables were expected to correlate with cohort since they were measured differently across cohort. These two variables were only included in CATSS specific analyses since the missingness in the NTR was high due to only including these questions in later surveys (see Table 3).

We also used Part 1 of the data to select the optimal set of meta-parameters of the different data mining methods such as the maximum number of trees necessary, number of splits for a given tree, shrinkage rate, and number of minimum observations in a terminal node in the tree methods. The motivation to train the model on a smaller portion of the data is to reduce the computational burden. The optimal tuning parameters are presented in Table 4.

Table 3. Imputation coefficients.

Variable	Imputed?	Mean	Total SE	W/in SE	B/w SE	Total Missing	NTR Missing	CATSS Missing
Aggression	No	0.2194	0.5780	0.5780	0.0058	0	0	0
Birth weight	Yes	-0.0184	0.9784	0.9783	0.0116	0.0435	0.0556	0.0286
Gest. Age	Yes	36.5213	2.6520	2.6518	0.0305	0.106	0.0482	0.1772
Maternal Smoking	Yes	0.2113	0.4083	0.4083	0.0057	0.2785	0.0463	0.5642
Maternal Alc.	No	0.1801	0.3843	0.3842	0.0069	0.4746	0.0477	1
Height	Yes	126.4288	6.5617	6.5614	0.0665	0.263	0.4217	0.0677
Weight	Yes	25.4686	5.2574	5.2570	0.0629	0.269	0.41	0.0955
Asthma	Yes	0.1154	0.3195	0.3195	0.0033	0.1669	0.2869	0.0192
Eczema	Yes	0.1292	0.3355	0.3354	0.0043	0.1661	0.2871	0.0172
Medication	Yes	0.1246	0.3303	0.3303	0.0036	0.438	0.7692	0.0304
Siblings	No	0.9033	0.8999	0.8998	0.0167	0.5811	0.2408	1
Both parents	Yes	0.8555	0.3516	0.3516	0.0038	0.3419	0.6063	0.0164
Mother Age	Yes	31.0550	4.1657	4.1653	0.0591	0.2013	0.0173	0.4278
Father Age	Yes	33.4881	5.0243	5.0239	0.0624	0.2191	0.0341	0.4468
Parenting score	No	3.6493	0.3061	0.3060	0.0056	0.4934	0.8816	0.0157
Mother Edu.	Yes	3.1880	0.7700	0.7699	0.0092	0.3667	0.2853	0.4668
Father Edu.	Yes	3.1356	0.8311	0.8310	0.0091	0.4151	0.3344	0.5143
Prop. LE	No	0.1119	0.1358	0.1358	0.0020	0.6155	0.756	0.4425
Motor skills	Yes	0.0269	0.1618	0.1618	0.0022	0.1869	0.2794	0.073
Argues	Yes	0.0662	0.2487	0.2487	0.0031	0.179	0.279	0.0559
Lying	Yes	0.0144	0.1191	0.1191	0.0015	0.1819	0.2798	0.0614
Stealing	Yes	0.0083	0.0909	0.0909	0.0011	0.1897	0.2776	0.0815
Brags	No	0.0585	0.2348	0.2348	0.0037	0.6033	0.281	1
Feels no guilt	No	0.9394	0.2386	0.2385	0.0044	0.6022	0.279	1
Short attention	Yes	0.0832	0.2761	0.2761	0.0035	0.2355	0.2879	0.1709
Daydream	Yes	0.0554	0.2289	0.2288	0.0026	0.2304	0.2776	0.1724
Distracted	Yes	0.0910	0.2876	0.2876	0.0035	0.4994	0.7675	0.1694
Doesn't finish things	Yes	0.0571	0.2321	0.2320	0.0032	0.4761	0.7677	0.1172
Hyperactive	Yes	0.0773	0.2670	0.2670	0.0040	0.2091	0.2774	0.1251
Superior	No	0.8962	0.3050	0.3050	0.0057	0.6021	0.2788	1
Bullied	Yes	0.0219	0.1462	0.1462	0.0015	0.18	0.2789	0.0583
Impulsive	No	0.0617	0.2406	0.2406	0.0040	0.603	0.2804	1
Fearful	Yes	0.0464	0.2104	0.2104	0.0027	0.4191	0.2862	0.5826
Anxious	Yes	0.0100	0.0997	0.0997	0.0013	0.4454	0.2785	0.6508
Nightmare	Yes	0.0159	0.1251	0.1251	0.0014	0.2199	0.2783	0.148

Note. Summary stats across imputed datasets for combined NTR and CATSS data in the exploratory partitions. Variables were excluded from imputation if the total missingness exceeded 50%, or if the variable was 100% missing in one of the two cohorts. Variability is expressed in the Total Standard Error (SE), which is comprised of SE for the estimate within each imputed dataset and variability between imputed datasets. Between SE is present due to missing data uncertainty. Gest. Age denotes gestational age; Alc. denotes alcohol use during pregnancy; Mother Edu. denotes maternal educational qualifications; Father Edu. denotes paternal educational qualifications; Prop. LE denotes the proportion of life events for which the response was yes; LE denotes Life Events. Note that small between-imputation variability is present for non-imputed variables because the partitions were randomized within imputed dataset.

Table 4. Optimal set of tuning parameters.

	Tree depth = 1	Tree depth = 5
Number of trees	18000	10000
Shrinkage	0.001	0.001
Minimum observations	100	100

Table 5. Average prediction error across imputed data sets for each data mining model.

Model	Average R^2	Average MSE
Hier_Lasso	0.219 (0.010)	0.260 (0.003)
Lasso	0.265 (0.009)	0.244 (0.003)
Stump	0.265 (0.009)	0.244 (0.003)
Interaction Tree	0.280 (0.009)	0.239 (0.003)
Augmented Stump	0.279 (0.009)	0.239 (0.003)

Note. Standard Deviations in parenthesis are for across imputed data sets.

Results Part 2 and Part 3 Data Analyses: Effect Type Selection and Predictor Selection

Effect Type Selection

Table 5 presents R^2 and MSE for the five models averaged across imputed data sets that were calculated in part 3 of the data. MSE measures the squared divergence of the model predicted outcome and the observed outcome aggregated over all subjects whereas R^2 expresses MSE as a proportion of the variance of the outcome. When computed in new data, prediction error quantifies how well a new observation would be predicted by the model while R^2 quantifies how well a model "explains" the individual differences in the outcome, in this case aggression.

The R^2 of the boosted interaction model was consistently higher than those of the other models; the MSE was consistently lower (see Table 5). These results provide evidence that there were at least some interactions and potentially non-linear effects. We then fit a boosted stump model to data that were augmented with pre-calculated product terms of the predictors. This model served to evaluate second order interaction effects, and is listed as augmented stump model in Table 5. The augmented stump model had an MSE and R^2 similar to the interaction tree model, permitting the conclusion that higher than second-order interaction effects did not contribute substantially to the prediction of physical/overt aggression. The augmented stump model was therefore used to select specific predictors and interaction effects to be included in the confirmatory prediction model fitted to Part 4 of the data.

Predictor Selection

We used Variable Importance Measures (VIMs) described by Friedman (2001) and integrated in the R package gbm to select predictors (Greenwell et al., 2019). The VIMs of a given model are scaled to sum up to 100, thus resulting in a percentage scale (i.e., percentage of the contribution of an individual predictor to the total prediction of a model). Since the objective of this study was to cast the net wide and investigate all possibly interesting predictors of aggression we used a cutoff of 0.5 to select predictors for the confirmatory last part of the analyses. This criterion resulted in selecting the following 12 main effects for the prediction model: argues (34.935), distracted (17.050), hyperactive (11.905), lying (4.023), stealing (1.927), daydreaming (1.549), being bullied (1.209), age mother (1.032), maternal smoking during pregnancy (0.957), sex of the child (0.836), education father (0.636), short attention (0.572), and living with both parents (0.519). In addition, the following 8 product terms had VIMs larger than 0.5: cohort x sex (3.874), cohort by daydreaming (3.330), argues x sex (2.653), argues x cohort (2.589), argues x distracted (1.208), stealing x cohort (1.153), hyperactive x sex (0.838), argues x age (0.557). As can be seen, with the exception of argues x distracted and argues x age the interaction terms involved either cohort or sex. These six interactions can be modeled using multi-group modeling while permitting group-specific regression coefficients. The two interactions involving arguing were investigated as product terms in the confirmatory model.

Results Part 4 Data Analysis: Multi-Group Regression Model

The effect type and predictor selection carried out in parts 2 and 3 of the data permitted a substantial reduction of possible effects from the initial 27 potentially interacting predictors to a total of 13 main and two interaction effects, and additionally six potentially group-specific effects.

We fit two multi-group regression models to the 35 imputed sets of Part 4 of the data, with sex and cohort as group defining variables. The base model featured the harmonized aggression outcome predicted by the 12 variables listed in the previous section and the two interaction terms (argues by distracted and argues by age), all with invariant regression coefficients across groups. Age was included as a main effect due to its participation in the interaction argues by age. This constrained model was compared to a model in which the variable argues was permitted to have sex and cohort specific effects, and the variable hyperactive to have sex specific effects. Group-specific regression coefficients capture the interactions with the grouping variable. The comparison of these two models evaluates the necessity of permitting interactions of arguing, daydreaming, stealing, and hyperactive with sex and/or cohort, respectively.

Comparing the two models in a likelihood ratio test aggregated over 35 imputed sets rejected the constrained model (chisq = 62.60, df = 6, p-value = 0). All other fit indices (AIC, BIC, RMSEA, and SRMR) were also consistently supporting the model with group-specific effects.

Table 6 shows the percentage out of 35 imputed sets each of the regression coefficients was significant. The table also shows the standardized regression coefficients using the group-specific variance of aggression for standardization. Note that standardized regression coefficients of dichotomous variables need to be interpreted with caution.

As can be seen in Table 6, the regression coefficients are in line with the ranking of VIMs in part 3 of the data, and reveal mild group differences. In general, arguing had the largest positive association with physical/overt aggression, followed by the items distracted and hyperactive. The variable daydreaming was significant in the CATSS sample in all 35 imputed sets but only in 31.4% in the NTR data, and had larger coefficients in CATSS. This difference is most likely due to item wording differences in the two cohorts (i.e., in CATSS but not in NTR the item wording included not listening when spoken to). Maternal smoking during pregnancy contributed to higher aggression, whereas living with both parents, higher educational level of the father, and higher age of the mother resulted in lower aggression. These effects were significant in almost all imputed sets, and can therefore be considered as robust. Short attention and the interactions of argues with distracted and with age were not significant in most imputed sets. Effect sizes can be found in Tables 6, 7 and 8, and are compared in detail to previous research in the discussion section.

Table 6. Part 4 analyses: Regression coefficients joint model/unequal coefficients.

	NTR male	NTR female	% of sets significant	CATSS male	CATSS female	% of sets significant
Maternal smoking	0.034	0.042	0.94	0.039	0.045	0.94
Both parents	-0.026	-0.031	0.91	-0.035	-0.040	0.91
Age mother	-0.029	-0.036	0.97	-0.034	-0.039	0.97
Education mother	-0.010	-0.013	0.43	-0.014	-0.016	0.43
Education father	-0.022	-0.027	0.97	-0.033	-0.039	0.97
Child argues	0.378	0.385	1.00	0.266	0.299	1.00
Child lies	0.079	0.072	1.00	0.066	0.069	1.00
Child steals	0.053	0.049	1.00	0.077	0.058	0.94
Child has short attention	0.007	0.006	0.09	0.007	0.006	0.09
Child daydreams	**0.021**	**0.020**	**0.31**	**0.086**	**0.069**	**1.00**
Child is distracted	0.138	0.118	1.00	0.157	0.136	1.00
Child is hyperactive	0.133	0.107	1.00	0.114	0.082	1.00
Child is being bullied	0.038	0.028	1.00	0.056	0.054	1.00
Arguing by distracted	-0.007	-0.006	0.03	-0.005	-0.004	0.03

Table 7. Part 4 analyses: Regression coefficients NTR-specific model.

	NTR male	NTR female	% of sets significant
Age of the child	0.040	0.050	0.97
Maternal alcohol	0.009	0.011	0.09
Child brags	0.127	0.067	1.00
Child feels no guilt	-0.035	-0.044	0.74
Child feels superior	-0.070	-0.082	1.00
Child is impulsive	0.099	0.089	1.00
Arguing by age	-0.228	-0.255	0.71

Table 8. Part 4 analyses: Regression coefficients CATSS-specific model.

	CATSS male	CATSS female	% of sets significant
Proportion life events	0.071	0.075	1.000
Parenting	-0.048	-0.054	0.914

The NTR-specific analyses revealed effects of bragging, impulsivity, and feelings of superiority (significant in all imputed sets, see Table 7). A small effect of age was significant in 97% of the imputed sets. Lack of guilt and the interaction arguing x age were significant in 74% and 71% of the imputed sets. Maternal alcohol consumption was only reaching statistical significance in 9% of the sets, thus not supporting a robust effect.

The CATSS specific analyses showed that the prevalence of serious life events was associated with an increased level of physical/overt aggression, whereas better parenting quality was associated with reduced aggression. These effects were significant in 100% and 91.4% of the data sets, respectively, and can therefore be considered as robust effects.

DISCUSSION

The present study modeled the relationship between childhood aggression and a wide range of predictor variables, using a novel methodological approach, in 62,227 children from two different cohorts. The large sample allowed for splitting the data in independent parts for exploration, variable selection, assessment of model performance, and fitting an interpretable confirmatory model. Employment of different data mining techniques provided the opportunity to investigate a large number of predictors simultaneously without the need to a priori specify which types of effects (i.e., linear or nonlinear main effects, linear or nonlinear interaction effects) were present in the data.

The most important variables were non-physical aggression (arguing), and two ADHD indicators (being easily distracted, and hyperactive), which were significant in 100% of the imputed sets. We report results concerning regression coefficients as "% of imputed sets" in which an effect was significant because this provides a measure of robustness of an effect whereas significance itself is less informative in large samples. Other variables that were significant predictors of higher childhood aggression in over 90% of the imputed sets were maternal smoking during pregnancy, the child not living with both parents, lower age of mother at birth, lower educational qualification of the father, lying, stealing, and being bullied. In addition, daydreaming was a significant predictor in the Swedish data. The cohort specific analyses revealed for the Dutch children that in more than 90% of the imputed sets, aggression was significantly predicted by older age, bragging, feeling superior, and impulsivity. For the Swedish sample, analyses revealed that a higher proportion of life events and lower levels of parenting (i.e., monitoring) were significantly associated with childhood aggression in more than 90% of the imputed sets. Variables with above-threshold VIMs that did not consistently predict childhood aggression across imputed data sets included lower educational qualification of the mother (significant in 43% of imputed data sets), short attention (significant in 9% of imputed data sets), an interaction between arguing and being easily distracted (significant in 3% of imputed data sets), and specifically for the Dutch children maternal alcohol use during pregnancy (significant in 9% of imputed data sets), feeling no guilt (significant in 74% of imputed data sets), and an interaction between arguing and age (i.e., the association between age and aggression varies between children who do and do not argue and vice versa; significant in 71% of imputed data sets). Variables not selected based on the zero or close to zero VIMs included birth weight, gestational age, height, weight, asthma, eczema, medication use, having siblings, age father at birth, motor skills, not finishing things, fear of situations, anxiety, and nightmares.

With regards to the demographic variables, sex and cohort interacted with some of the variables (e.g., argues x sex, argues x cohort, hyperactive x sex), which led to multi-group models. Age only appeared to have an effect for the Dutch children, with higher aggression for older children. The group differences for boys and girls, and for the Netherlands and Sweden provide evidence for etiological differences between these groups.

Of the prenatal characteristics, maternal smoking during pregnancy significantly predicted childhood aggression in 94% of the imputed sets; maternal alcohol use during pregnancy was only significant in 9% of the imputed sets (only measured in Dutch children), and birth weight and gestational age were not selected based on their VIMs. Possibly, the influence of birth weight and gestational age on childhood aggression are attenuated by the environment in which children grow up (LaPrairie, Schechter, Robinson, & Brennan, 2011). The effect of .009 to .011 of maternal alcohol use during pregnancy was similar to the correlation of .008 found in a sample from the United Kingdom (Malanchini et al., 2018). Maternal smoking during pregnancy had an effect on childhood aggression ranging between .034 and .045, which was slightly smaller than the correlation of .085 found in a recent meta-analysis, with partly overlapping samples (Malanchini et al., 2018). Differences in effect sizes between previous research and the present study could be explained by the fact that we investigated all predictors simultaneously, which avoids overestimation of correlated effects.

None of the physical development variables were selected based on their importance (i.e., height, weight, asthma, and medication use). A meta-analysis on the association between asthma and externalizing behavior revealed an association of .29 (Pinquart & Shen, 2011). In addition, previous research found associations between height and weight during early childhood and later aggressive behaviors (i.e., d = 0.25 – 0.30), but when controlling for other factors such as socioeconomic status, the associations disappeared (Raine, Reynolds, & Venables, 1998). Overall, it appears that the association between physical development and childhood aggression might be overruled when taking measures of the family environment and behavioral symptoms into account.

From the family environment variables (i.e., siblings, whether both parents live in the same household, age mother at birth, and age father at birth), only whether both parents lived in the same household (i.e., effects from -.040 to -.026), and age of mother at birth (i.e., effects from -.039 to -.029) were included in the confirmatory model. That having siblings was not a strong predictor for childhood aggression might be explained by the fact that all the children in the samples were twins, which may attenuate the impact of having other siblings. A possible explanation for higher aggression when parents do not live in the same household could be an increase in parental stress due to single parenthood, such as a lowered income (Briggs, Cox, Sharkey, Briggs, & Black, 2016). The positive effect of higher maternal age was in line with previous research (Tearne et al., 2015). It could be due to older mothers having better socioeconomic circumstances (Bornstein, Putnick, Suwalsky, & Gini, 2006), higher satisfaction with parenting, and more time spent with children (Ragozin, Basham, Crnic, Greenberg, & Robinson, 1982).

The parenting variable pertained to parental monitoring, an established predictor for childhood aggression (Racz & McMahon, 2011). In the present study, the regression coefficients (only measured in Swedish children) were -.048 and -.054, indicating higher aggression for children whose parents monitor them less. Previous research found an association of .31 between poor supervision and oppositional defiant disorder (ODD) and .39 between poor supervision and conduct disorder (CD; Burke, Pardini, & Loeber, 2008), which was stronger than the regression coefficients in the present study. Both disorders co-occur with aggressive behaviors, together

with symptoms that were important predictors in the present study, namely arguing for ODD, and stealing and lying for CD. Possibly, the regression coefficients for parenting were diminished due to inclusion of these variables in the confirmatory model.

Both maternal and paternal education were included in the confirmatory model. Paternal education was significant in 97% of the imputed sets, maternal education only in 43%. Both indicated lower levels of aggression for children when parents are more highly educated. The effects for paternal education ranged from -.039 to -.022, for maternal education they ranged from -.016 to -.010; these estimates were close to the correlation found of -.099 between aggression and socioeconomic status (which can be assessed through parental education level; Winkleby, Jatulis, Frank, & Fortmann, 1992) in a meta-analysis (Piotrowska et al., 2015)

A higher proportion of life events predicted higher aggression with regression coefficients of .071 and .075. Previous research found associations of .16 and .28 between life events and aggression (Guerra et al., 1995; Mcknight, Huebner, & Suldo, 2002). The discrepancy between the coefficients in the present study and previous research could be caused by heightened behavioral symptoms as a result of exposure to life events and thus absorbing the effects.

The importance of the mother-reported behavioral symptoms is in line with the high comorbidity between childhood aggression and other behavior problems (Bartels et al., 2018). The most important predictors were symptoms related to attention-deficit hyperactivity disorder (ADHD; i.e., hyperactivity, being easily distracted, and impulsivity), oppositional defiant disorder (ODD; i.e., arguing), and conduct disorder (CD; i.e., lying, stealing). The importance of arguing (i.e., regression coefficients between .266 and .385), lying (i.e., regression coefficients between .066 and .079), and stealing (i.e., regression coefficients between .049 and .077) confirms the overlap between childhood aggression and ODD and CD (American Pyschiatric Association, 1994), although the symptoms reflect distinguishable constructs in 9-year-old children (Lubke, McArtor, Boomsma, & Bartels, 2018). In addition, ADHD often co-occurs with aggressive behavior (e.g., Harvey, Breaux, & Lugo-Candelas, 2016; Rhee, Willcutt, Hartman, Pennington, & DeFries, 2008), which was supported by the regression coefficients in this study for being easily distracted (i.e., .118 - .157), hyperactivity (i.e., .082 - .133), impulsivity (i.e., .089 - .099), and daydreaming (i.e., 0.20 – 0.21 in the Dutch sample and 0.69 – 0.86 in the Swedish sample). This could partially be explained by a shared genetic liability; research found high genetic correlations (e.g., .46 - .74) between ADHD behaviors and forms of childhood aggression (Dick, Viken, Kaprio, Pulkkinen, & Rose, 2005; Kuja-Halkola, Lichtenstein, D'Onofrio, & Larsson, 2015). Previous research finding no association between short attention and aggression, was confirmed by the fact that, in the present study, short attention was only significant in 9% of the imputed sets (i.e., regression coefficients of .006-.007; e.g., Nagin & Tremblay, 2001).

Feeling no guilt, bragging, and feeling superior are all related to psychopathic traits including callous/unemotional (CU) traits and narcissism (Salekin, 2017). In the present study, these were all significant predictors for childhood aggression (i.e., respectively, regression coefficients of -.035, .127, and -.070 for boys and -.044, .067, and -.082 for girls). Taking the direction of item wording into account, this indicates higher aggression when children feel no guilt, brag, and feel superior, in line with previous research (Kerig & Stellwagen, 2010; Svensson et al., 2018). Being bullied had an effect between .028 and .056, with higher aggression for children being bullied. These effects were slightly smaller than a correlation of .14 found in a meta-analysis (Reijntjes et al., 2011). Motor skills were not selected based on VIMs, which was in line with previous research on behavior problems and motor skills mainly finding effects for ADHD, and thus not aggression (Emck, Bosscher, Beek, & Doreleijers, 2009). Internalizing symptoms (i.e., fear of situations, anxiety, nightmares) were not selected based on their VIMs. This could be explained by the lower comorbidity between aggression and internalizing disorders compared to the comorbidity between aggression and externalizing disorders (Bartels et al., 2018), indicating that internalizing symptoms are less important in predicting childhood aggression compared to externalizing symptoms.

Taking many variables into account simultaneously could explain that most of the regression coefficients in the present study were smaller than reported in previous research. While the findings are correlational, and should thus not be interpreted as causal relations, they do provide direction for variables that are valuable to examine through longitudinal research. The prediction effect of externalizing behavior symptoms remained taking all other selected variables into account. Practically, this implies that, although environmental variables may be important for the development of childhood aggression, paying attention to behaviors of the child such as arguing, being easily distracted, and hyperactivity will yield a better prediction of childhood aggression.

Heterotypic continuity causes childhood aggression to express differently across the life span (e.g., Hannigan, Walaker, Waszczuk, McAdams, & Eley, 2017; Lubke, McArtor, Boomsma, & Bartels, 2018), which hampers effective diagnosis and treatment referral. The high co-occurrence of childhood aggression and other symptoms suggests that having elevated levels for one disorder will also imply elevated levels for other disorders (Bartels et al., 2018). Screening for behavioral symptoms that underlie multiple mental disorders may target children that will likely develop childhood aggression but could also develop another disorder, such as ADHD. This may aid early detection of a liability to develop psychopathology and implementation of treatment before children develop a full-blown disorder. Moreover, treating one disorder could also positively affect levels for other disorders. For example, research found improved levels of aggression as a result of treatment

for ADHD (Chan, Fogler, & Hammerness, 2016). Our results, together with previous research, suggest merit in monitoring children's behavioral symptoms, because they might predict later aggression or any other disorder that often co-occurs with childhood aggression.

Although we included a wide range of predictor variables, some known risk factors for childhood aggression were not collected by the included cohorts or not available for the present study, imposing a limitation on the comprehensiveness of the prediction model. Examples are exposure to domestic violence (Evans, Davies, & DiLillo, 2008) and parental psychopathology (Connell & Goodman, 2002; Goodman et al., 2011). Nevertheless, these unavailable variables apply to more extreme cases, whereas the predictors in the present study would apply to the general population. Moreover, it may be easier to observe the variables included in the present study than to obtain information on sensitive topics. The predictors that were most important in the present study comprise salient behaviors that could be noticed by, for instance, parents or teachers, making the findings feasible to apply in common practices.

Throughout the study, we applied rigorous and novel methodological approaches. First of all, partitioning the data into four independent parts provided us with the possibility to fit different models with different types of predictor effects (i.e., linear, nonlinear, interaction) without having to pre-specify them. Moreover, partitioning the data allowed for testing the models in independent sets of data, thereby preventing the risk of overfitting. Second, analyses were able to detect measurement non-invariance variance related to differences between cohorts. For example, the behavioral item on daydreaming did not come up in the EDA as different between cohorts, but the data mining analyses followed by confirmatory modeling revealed that there was in fact measurement non-invariance. Therefore, we are confident that with these rigorous methods and the data available to us, we have obtained the most robust prediction model for childhood aggression.

Childhood aggression is a very heterogeneous disorder (Bolhuis et al., 2017). This may explain that research so far is inconclusive on the etiology of childhood aggression and that treatment effectiveness for childhood aggression is still limited (Hendriks et al., 2018; Weisz et al., 2017). Therefore, it is important to clearly define the type of childhood aggression under scrutiny (Hofvander, Ossowski, Lundström, & Anckarsäter, 2009). In the present study, we accounted for this by specifically examining physical and overt aggression. For future research, it would be interesting to study whether our findings also apply to other types of childhood aggression.

Finally, childhood aggression is a developmental disorder with a strong genetic component (e.g., 32% - 83%; Hudziak et al., 2003; Porsch et al., 2016; Rhee & Waldman, 2002). Moreover, many of the predictor variables included in the present study are partly explained by genetic factors such as parenting and being bullied in secondary school (Veldkamp et al., submitted; Vinkhuyzen, Van Der Sluis, De Geus, Boomsma, & Posthuma, 2010), and thus possibly even overlap in genetic liability, which may have led to biased estimates. One way to take this into account would be to correct for genetic information, for instance by including a polygenic risk score (e.g., Wray, Goddard, & Visscher, 2007) for childhood aggression or strongly related variables in the prediction model. The currently available GWAS for childhood aggression (Pappa et al., 2016) and antisocial behavior (Tielbeek et al., 2017) were not sufficiently powered to use for a polygenic risk score. Nevertheless, this will likely become possible in the near future, which then will provide opportunities to obtain a clearer understanding of the relationship between childhood aggression and predictor variables.

In conclusion, the presented research is the first large-scale study that included a large number of potential predictors for childhood aggression. The large number of variables allows one to assess the presence of all possible main and interaction effects simultaneously. Effects were detected using deductive data mining, and were tested using a confirmatory model fitted to a holdout partition of the data. Investigating multiple predictors simultaneously results in more unbiased effect sizes compared to one-at-a-time analyses, and form a more reliable basis for future research into the prediction of childhood aggression. The most important predictors were salient behaviors such as arguing, being easily distracted, and hyperactivity. Recommendations for future research include testing the found relations in longitudinal data to establish direction of causality and adding genetic information to control for genetic overlap between variables in the prediction model. Altogether, the present study applied rigorous methods on a wide range of predictor variables and yielded a set of variables that may facilitate early detection and prevention of childhood aggression.

Supplementary Table 1. Overt/Physical Aggression Items in ACTION

Item Code	Item
A-TAC63	Has there ever been a time when he/she would be angry to the extent that he/she cannot be reached?
A-TAC65*	Does he/she often tease others by deliberately doing things that are perceived as provocative?
A-TAC70	Has he/she ever been deliberately been physical cruel to anybody?
A-TAC71	Does he/she often get into fights?
CBCL016	Cruelty, bullying or meanness to others
CBCL020	Destroys his/her own things
CBCL021	Destroys things belonging to his/her family or others
CBCL023	Disobedient at school
CBCL037	Gets in many fights
CBCL057	Physically attacks people
CBCL094	Teases a lot
CBCL095	Temper tantrums or hot temper
SDQ05	Often has temper tantrums or hot tempers
SDQ07	Generally obedient, usually does what adults request
SDQ12	Often fights with other children or bullies them
MPNI13	Teases other kids or attacks them for no reason at all
MPNI21	Hurts other kids when angry, e.g. by hitting, kicking, or throwing things at them
MPNI25	Bullies smaller and weaker kids
MPNI27	Calls people names when angry at them
MPNI33	Is disobedient at school/home

Supplementary Table 2. Variables per category and their description

Variable category	Variable coding	CATSS	NTR
Outcome variable	Harmonized aggression factor score	x	x
Demographics	Sex	x	x
	Age	x	x
	Cohort	x	x
Prenatal characteristics	Birth weight, standardized separately per cohort	x	x
	Gestational age, in weeks	x	x
	Maternal smoking during pregnancy; yes/no	x	x
	Maternal alcohol use during pregnancy; yes/no	-	x
Physical development	Height, in cm, parent report. This variable was corrected for age differences.	x	x
	Weight, in kilos, parent report. This variable was corrected for age differences.	x	x
	Asthma, parent report; yes /no	Have or had ... Asthma	Asthma, chronic bronchitis or CARA ...,
	Eczema, parent report ; yes/no	Have or had ... Eczema	Serious skin disease or eczema
	Medicine use, parent report; yes/no	Does s/he use prescribed medicine?	Does the child currently use prescription medication?
Family environment	Siblings, other than the co-twin, parent report; yes/no	-	x
	Whether both parents live in the same household, parent report; yes/no	x	x
	Age mother at birth	x	x
	Age father at birth	x	x

Supplementary Table 2. Continued

Parenting	Parenting, reported by parents. Mean score of >66% of the items, ranging from 1-4. A higher score indicates more parental monitoring	Do you know what your child does during his/her free time? Do you usually know what kind of homework your child has? Does your child keep a lot of secrets from you about what s/he does during her/his free time? RECODED Does your child need to have your permission to stay out late on weekday evening? Do you always require that your child tells you where s/he is at night, who s/he is with and what they do together? How often do you initiate a conversation with your child about things that happened during normal day at school? Do you usually ask your child to talk about things that happened during her/his free time (whom s/he met when s/he was out in the city, free time activities, etc.)? Does your child usually want to tell how school was when s/he gets home (how s/he did on different exams, her/his relationship with teachers, etc.)? In original data coded 1, 2, 3, 4, 5, for data aligning 3 and 4 combined.	I always know where and with whom my child is if he - she is not at home I know the interest of the child and what kind of hobbies he -she likes I know the daily activities of the child I take into account to the view of the child
Parental education level	Maternal education level, 1-4: 1 = elementary school; 2 = high school, but not finished; 3 = finished high school; 4 = university.	x	x
	Paternal education level, 1-4: 1 = elementary school; 2 = high school, but not finished; 3 = finished high school; 4 = university.	x	x

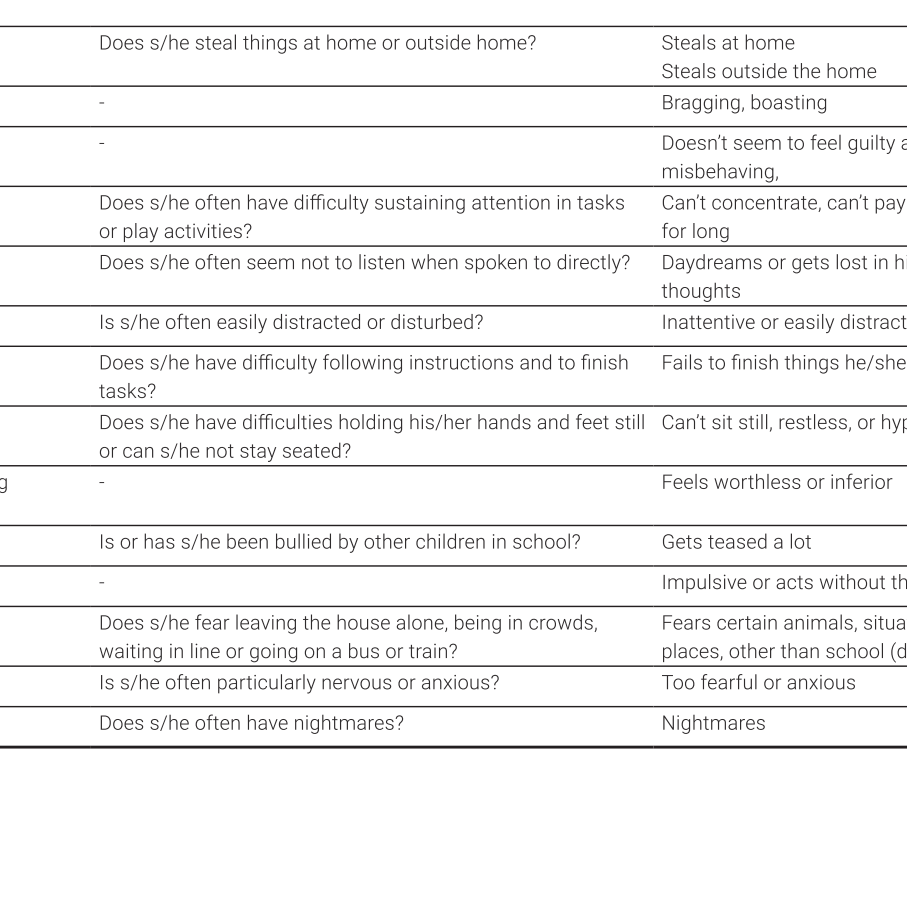
Supplementary Table 2. Continued

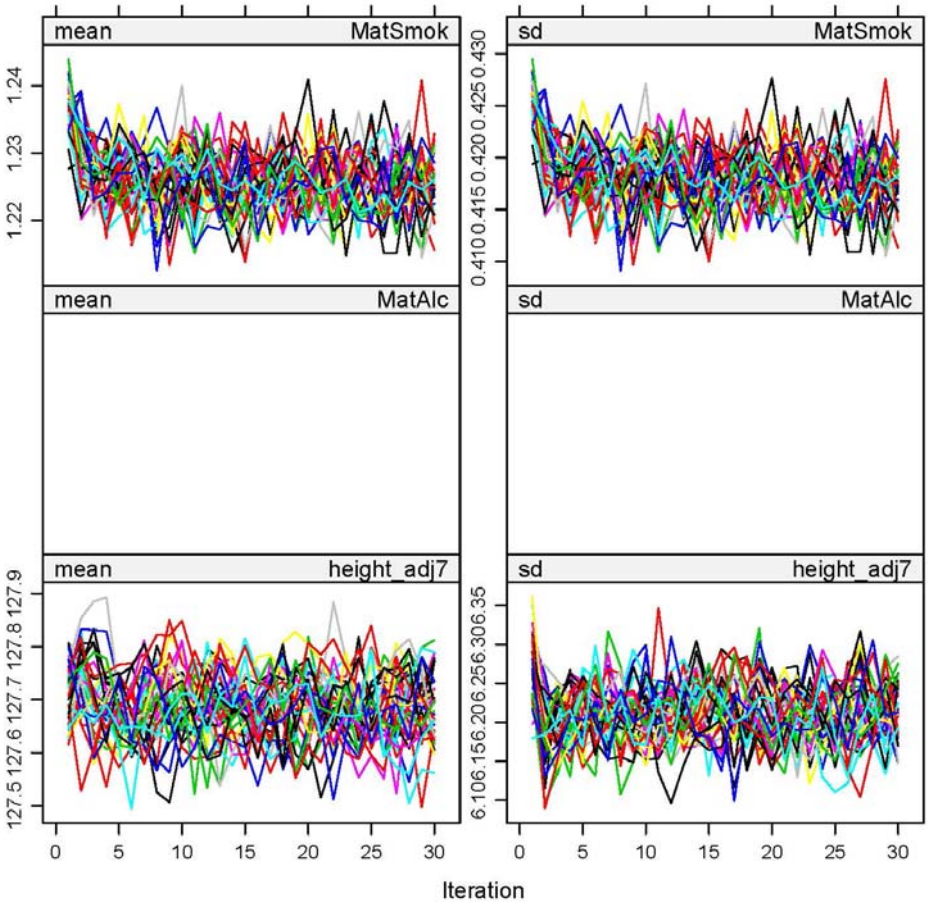
Life events	Proportion of life events, calculated for all observations with more than 66% of the items filled in. Parent report for CATSS, self-report for NTR.	Parent-reported. Has the child ever been in a serious car accident where he/she got modest to severe physical injuries or needed medical care? Has the child ever been in any other serious accident where he/she got injured or hospitalized? Has the child ever been emotionally abused or neglected? For example, being frequently shamed, embarrassed, ignored, or repeatedly told that he/she were "no good" Has the child ever been physically neglected? For example, not fed, not properly clothed, or left to take care of him/herself Has the child ever been physically abused – for example, hit, choked, burned, or beaten or severely punished by someone he/she knew well? Was the child ever touched or made to touch someone else in a sexual way, because he/she felt forced in some way or threatened by harm to him/herself or someone else ? Has the child ever had sex because he/she felt forced in some way or threatened by harm to him/herself or someone else? With sex, we mean orally, anally, and/or genitally '. Has the child ever observed physical violence between family members? For example, hitting, kicking or punching Has the child ever witnessed a threatening or violent criminal incident? (not film, internet or TV) e.g. someone was attacked, seriously hit, robbed or stabbed Has the child ever been a direct witness (not film, Internet or TV) to any other serious incident that you have not mentioned? Is there any other cruel or terrifying event the child has been exposed to?	Self-reported. You moved to another neighborhood or city A good friend moved house You changed schools (not from elementary to high school) You were seriously ill or had a serious accident Someone close to you was or is seriously ill Someone close to you died Your parents have serious conflicts - fights Your mother or father left home or your parents got divorced Your mother's or father's new partner came to live with you Your brother or sister left home Your mother or father became unemployed Your mother or father started working again after a long period at home A little brother or sister was born or adopted
Behavioral symptoms mother report	Motor skills; 0-1	Does s/he have problems coordinating movements smoothly?	Poorly coordinated or clumsy
	Arguing; 0-1	Does s/he often argue with adults?	Argues a lot
	Lying; 0-1	Does s/he often lie or cheat?	Lying or cheating

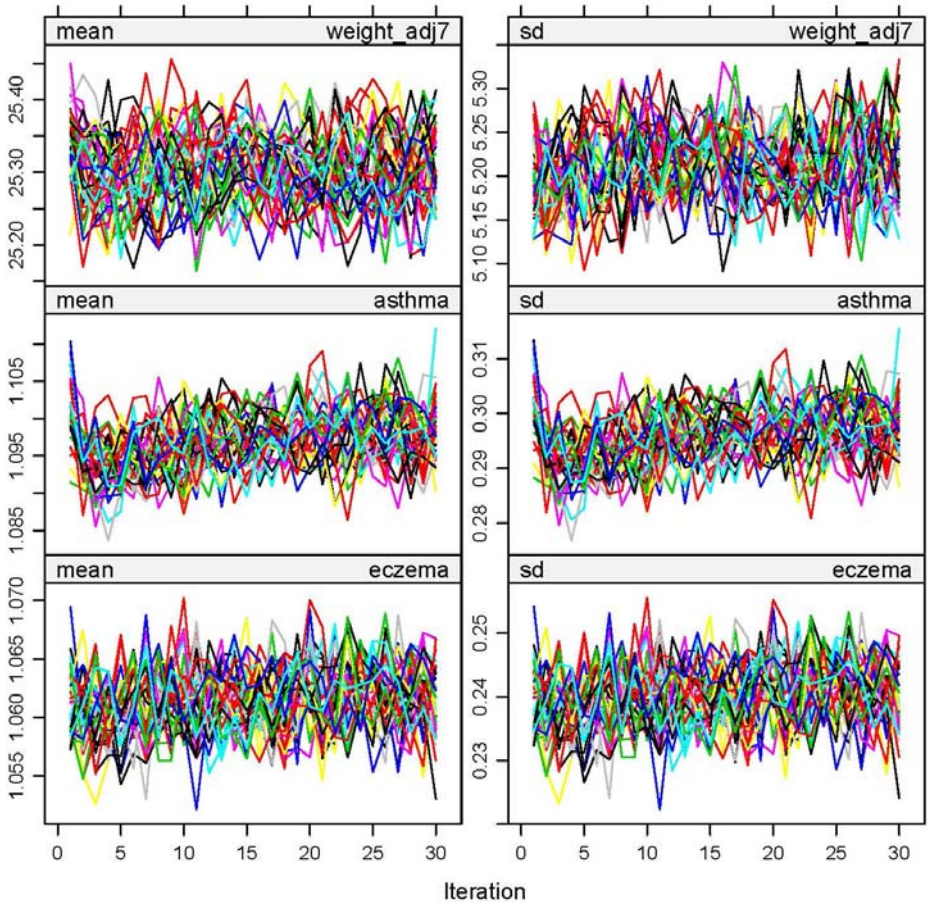
Supplementary Table 2. Continued

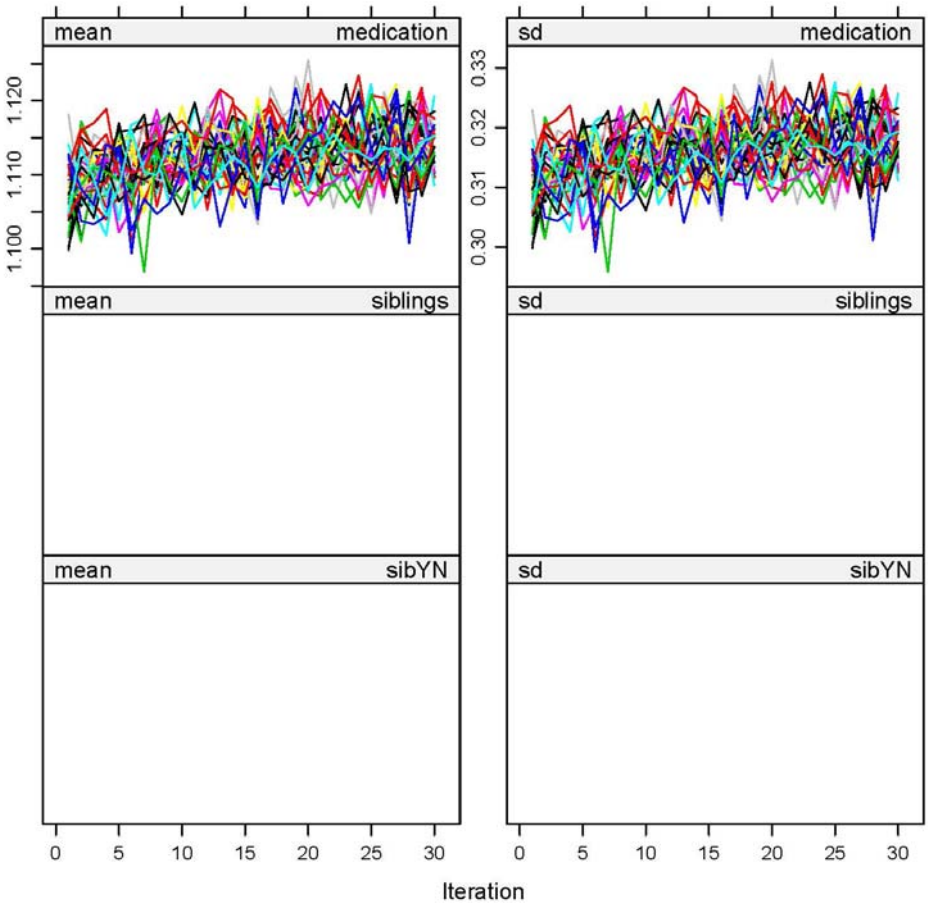
	Stealing; 0-1	Does s/he steal things at home or outside home?	Steals at home Steals outside the home
	Bragging; 0-1	-	Bragging, boasting
	No guilt; 0-1, 0 indicates less guilt	-	Doesn't seem to feel guilty after misbehaving,
	Short attention; 0-1	Does s/he often have difficulty sustaining attention in tasks or play activities?	Can't concentrate, can't pay attention for long
	Daydreaming; 0-1	Does s/he often seem not to listen when spoken to directly?	Daydreams or gets lost in his/her thoughts
	Easily distracted; 0-1	Is s/he often easily distracted or disturbed?	Inattentive or easily distracted
	Not finishing things; 0-1	Does s/he have difficulty following instructions and to finish tasks?	Fails to finish things he/she starts
	Hyperactivity; 0-1	Does s/he have difficulties holding his/her hands and feet still or can s/he not stay seated?	Can't sit still, restless, or hyperactive
	Feeling superior; 0-1, 0 indicates feeling superior	-	Feels worthless or inferior
	Being bullied; 0-1	Is or has s/he been bullied by other children in school?	Gets teased a lot
	Impulsivity; 0-1	-	Impulsive or acts without thinking
	Fear of situations; 0-1	Does s/he fear leaving the house alone, being in crowds, waiting in line or going on a bus or train?	Fears certain animals, situations, or places, other than school (describe):
	Anxiety; 0-1	Is s/he often particularly nervous or anxious?	Too fearful or anxious
	Nightmares; 0-1	Does s/he often have nightmares?	Nightmares

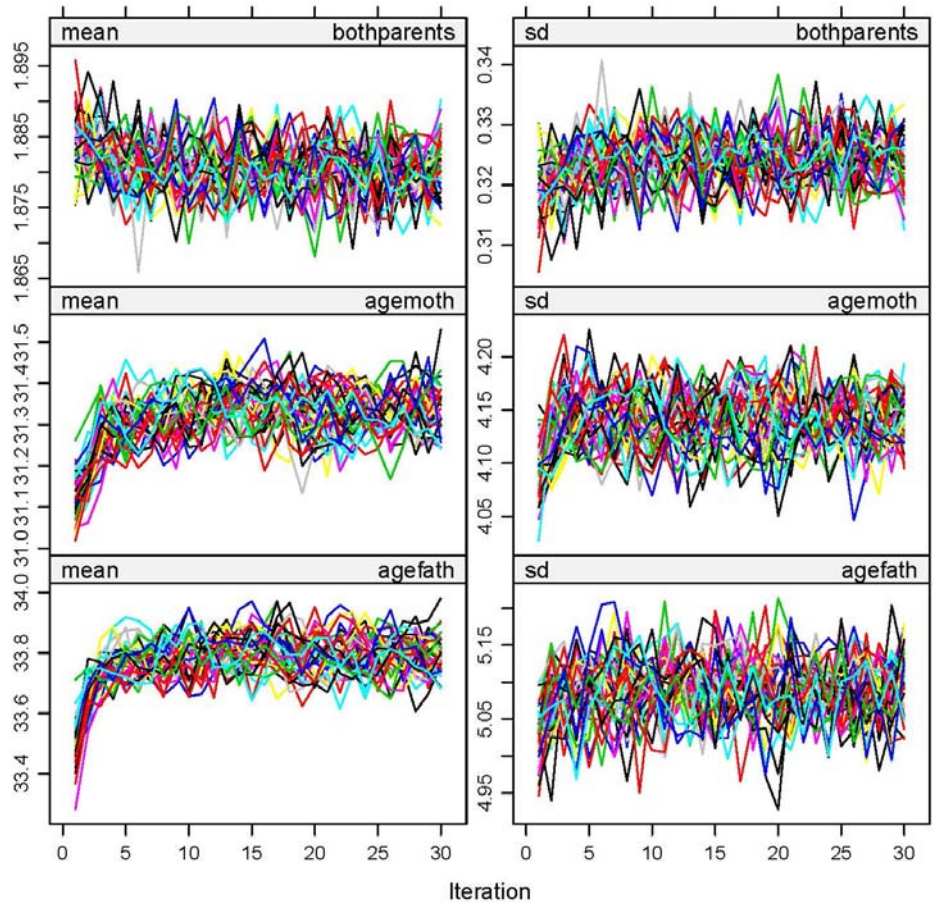
Supplementary Figure 1. Imputation convergence plots

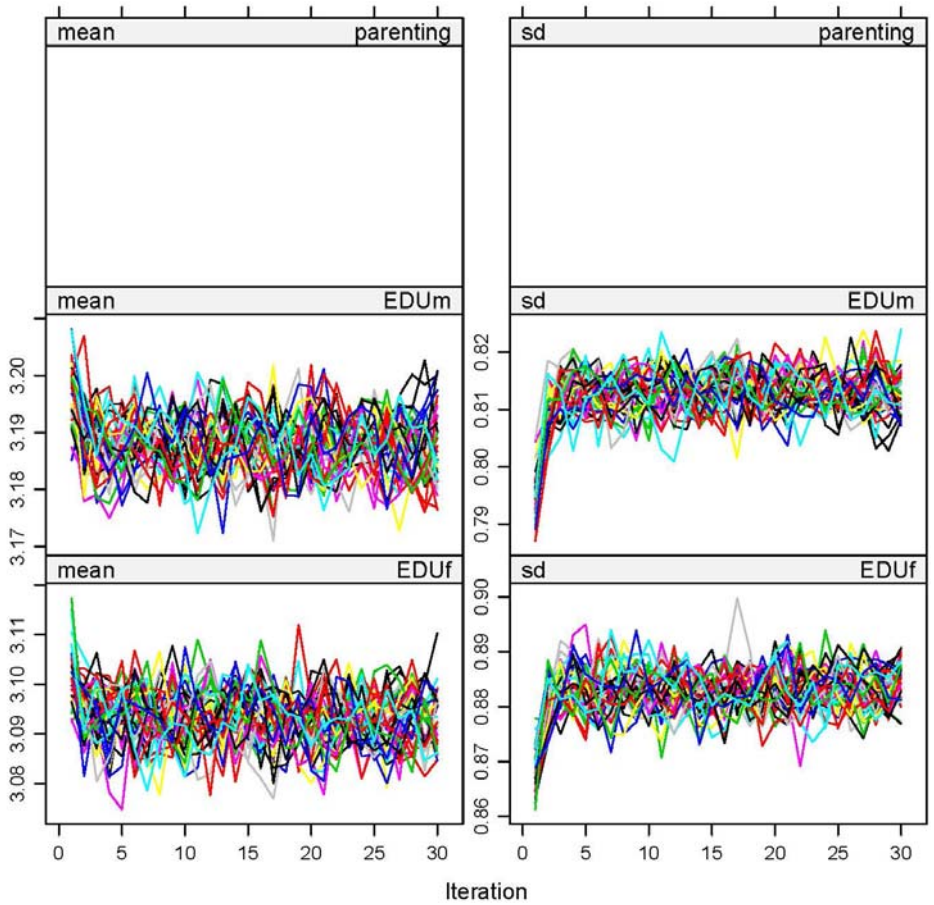


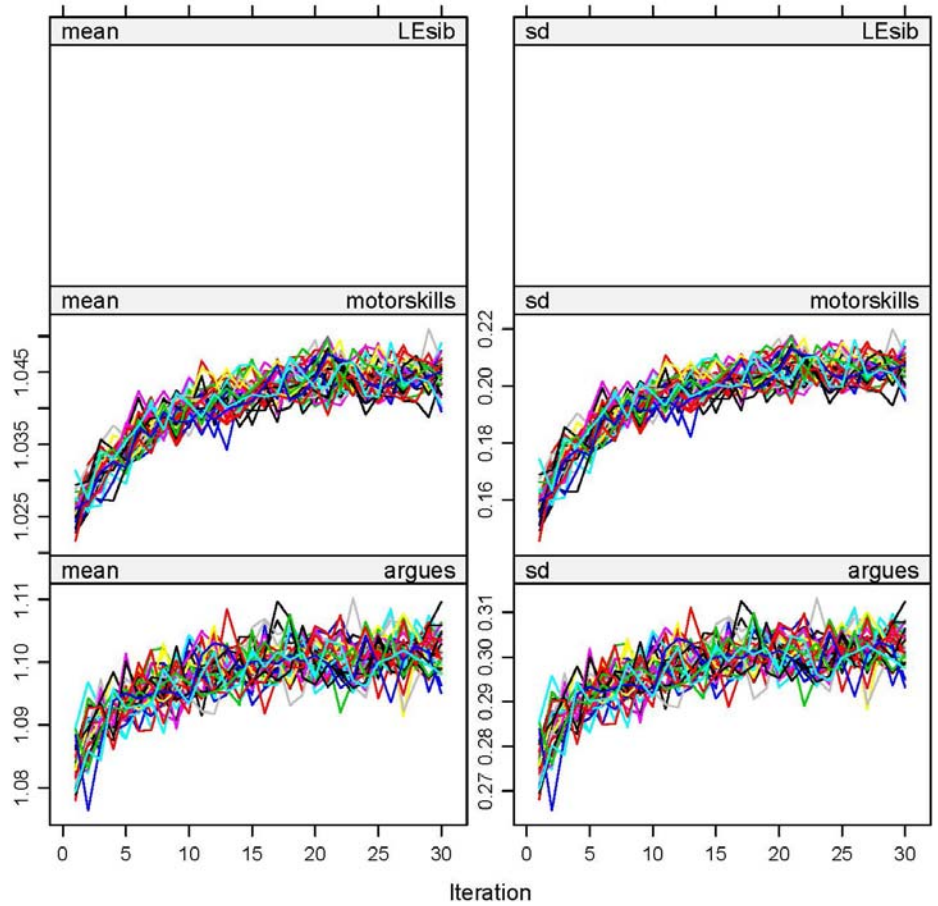

mean
MatSmok
sd
MatSmok
mean
MatAlc
sd
MatAlc
mean
height_adj7
sd
height_adj7
1.22
1.23
1.24
0.410
0.415
0.420
0.425
0.430
127.5
127.6
127.7
127.8
127.9
6.10
6.15
6.20
6.25
6.30
6.35
0
5
10
15
20
25
30
Iteration

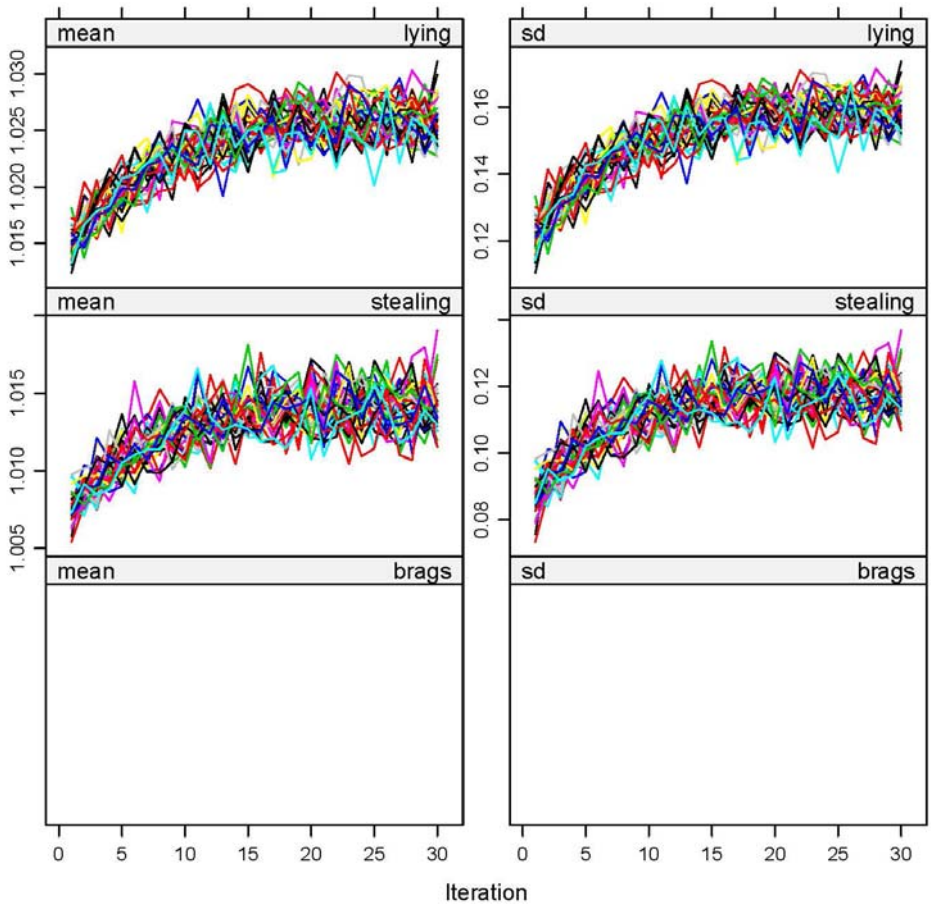

mean
weight_adj7
sd
weight_adj7
mean
asthma
sd
asthma
mean
eczema
sd
eczema
25.20
25.30
25.40
5.10
5.15
5.20
5.25
5.30
1.085
1.095
1.105
0.28
0.29
0.30
0.31
1.055
1.060
1.065
1.070
0.23
0.24
0.25
0
5
10
15
20
25
30
Iteration

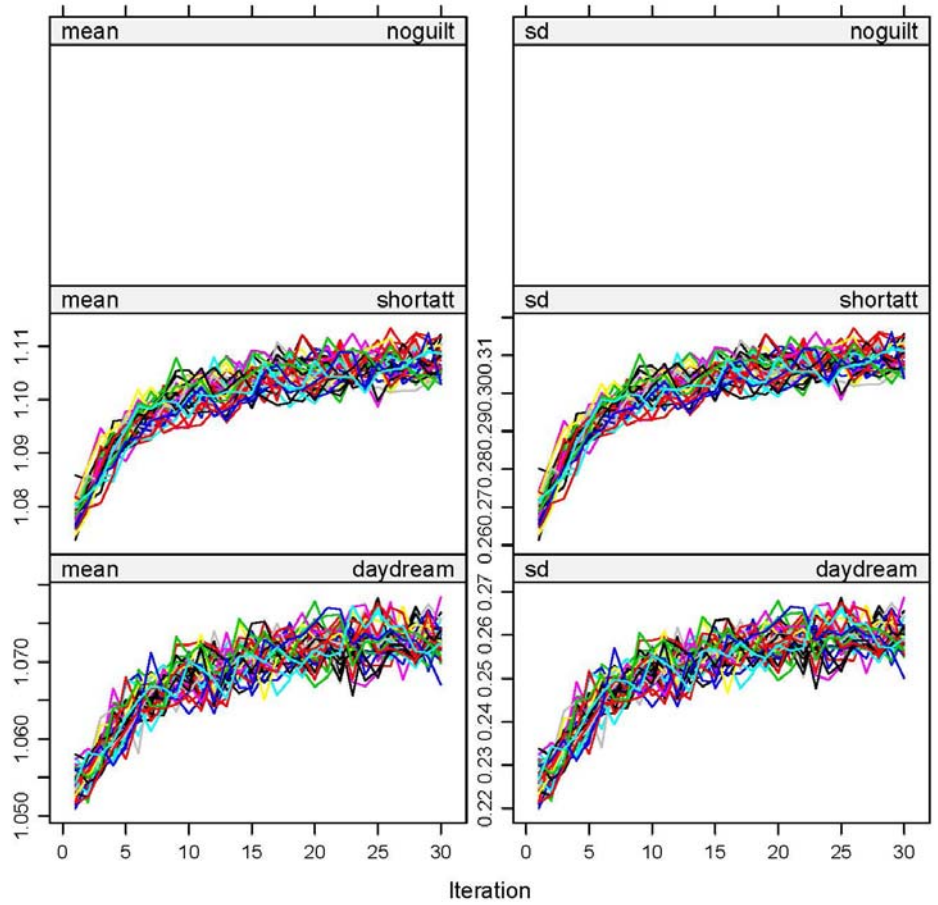

mean
medication
sd
medication
mean
siblings
sd
siblings
mean
sibYN
sd
sibYN
1.100
1.110
1.120
0.30
0.31
0.32
0.33
0
5
10
15
20
25
30
Iteration

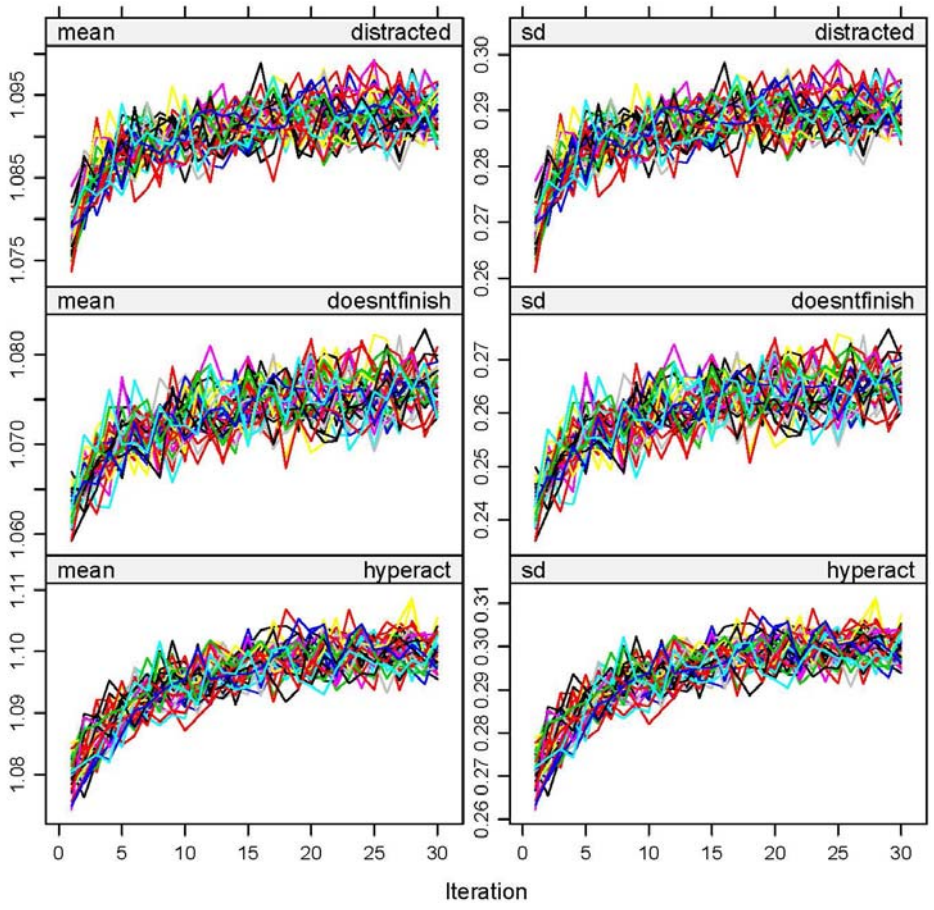

mean
bothparents
sd
bothparents
mean
agemoth
sd
agemoth
mean
agefath
sd
agefath
1.865
1.875
1.885
1.895
0.31
0.32
0.33
0.34
31.0
31.1
31.2
31.3
31.4
31.5
4.05
4.10
4.15
4.20
33.4
33.6
33.8
34.0
4.95
5.05
5.15
0
5
10
15
20
25
30
Iteration

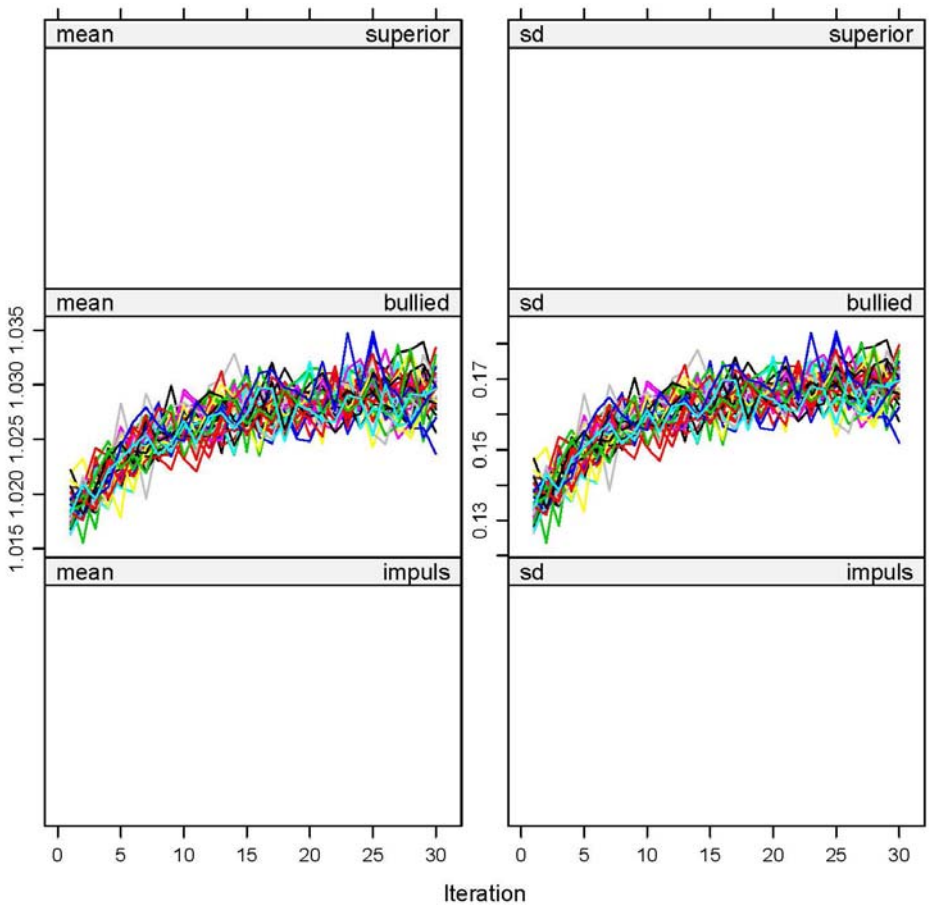

mean
parenting
sd
parenting
mean
EDUm
sd
EDUm
mean
EDUf
sd
EDUf
3.20
3.19
3.18
3.17
3.11
3.10
3.09
3.08
0.82
0.81
0.80
0.79
0.90
0.89
0.88
0.87
0.86
0
5
10
15
20
25
30
Iteration

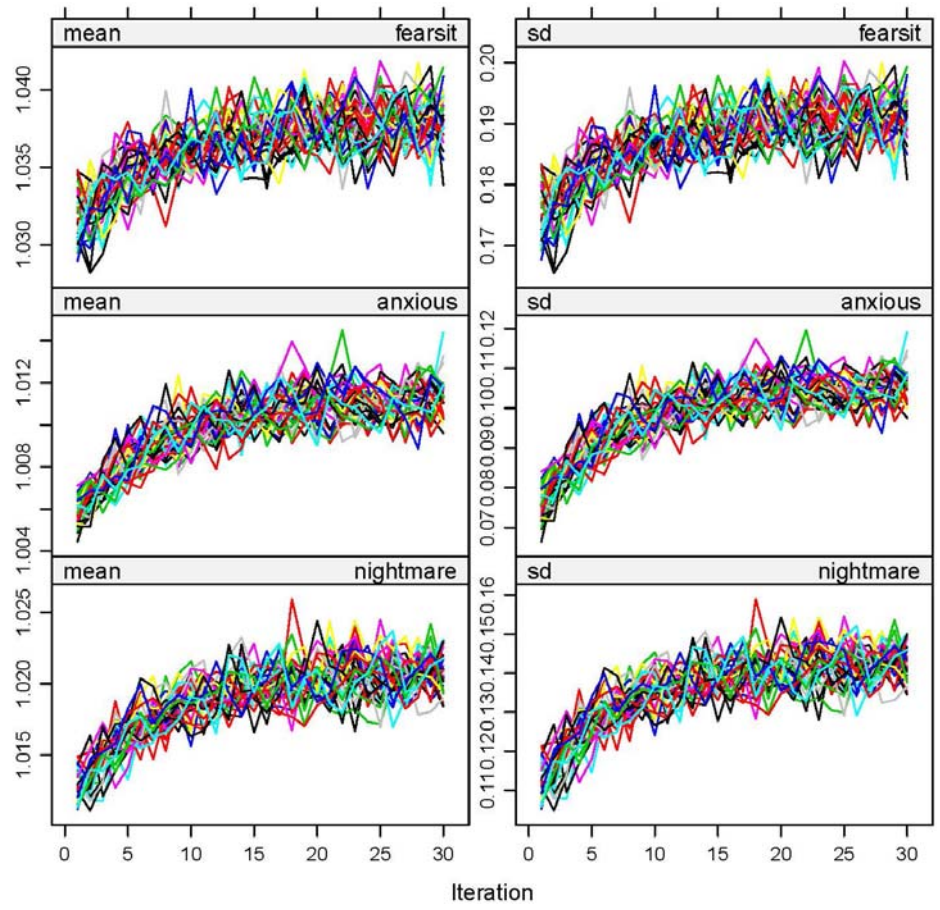

mean
LEsib
sd
LEsib
mean
motorskills
sd
motorskills
mean
argues
sd
argues
1.045
1.035
1.025
1.11
1.10
1.09
1.08
0.22
0.20
0.18
0.16
0.31
0.30
0.29
0.28
0.27
0
5
10
15
20
25
30
Iteration


mean
lying
sd
lying
mean
stealing
sd
stealing
mean
brags
sd
brags
1.005 1.010 1.015
1.015 1.020 1.025 1.030
0.08 0.10 0.12
0.12 0.14 0.16
0 5 10 15 20 25 30
Iteration


mean
noguilt
sd
noguilt
mean
shortatt
sd
shortatt
mean
daydream
sd
daydream
1.08 1.09 1.10 1.11
1.050 1.060 1.070
0.22 0.23 0.24 0.25 0.26 0.27
0.260.270.280.290.300.31
0 5 10 15 20 25 30
Iteration


mean
distracted
sd
distracted
mean
doesntfinish
sd
doesntfinish
mean
hyperact
sd
hyperact
1.075
1.085
1.095
1.060
1.070
1.080
1.08
1.09
1.10
1.11
0.26
0.27
0.28
0.29
0.30
0.24
0.25
0.26
0.27
0.26
0.27
0.28
0.29
0.30
0.31
0
5
10
15
20
25
30
Iteration


mean
superior
sd
superior
mean
bullied
sd
bullied
mean
impuls
sd
impuls
1.015
1.020
1.025
1.030
1.035
0.13
0.15
0.17
0
5
10
15
20
25
30
Iteration


mean
fearsit
sd
fearsit
mean
anxious
sd
anxious
mean
nightmare
sd
nightmare
1.040
1.035
1.030
0.20
0.19
0.18
0.17
1.012
1.008
1.004
0.12
0.11
0.10
0.09
0.08
0.07
1.025
1.020
1.015
0.16
0.15
0.14
0.13
0.12
0.11
0
5
10
15
20
25
30
Iteration